新闻数据可视化工具“词云”在新媒体中的创新实践

秦玉芳 黎若楠 刘颖旭
（新华社通信技术局，北京 100803）

摘 要：随着互联网技术和新媒体业务的蓬勃发展，数据新闻创新产品层出不穷，新闻的数据可视化成为一种趋势，国内外的数据可视化实践也正在如火如荼的进行。在众多的新闻可视化形式中，词云可以快速概览新闻内容并获取关键信息，且呈现形式多样，受到了编辑记者和读者的喜爱。本文以新闻数据可视化工具“词云”为切入点，基于对新闻报道的场景需求，利用新闻关键词提取算法和数据可视化技术实现词云工具，有效地提升了新华社编辑记者制作词云的效率。

关键词：互联网技术；词云；关键词提取；数据新闻；可视化；数据 **中图分类号：**TN816 **文献标识码：**A
文章编号：1671-0134（2021）09-046-04 DOI：10.19483/j.cnki.11-4653/n.2021.09.013
本文著录格式：秦玉芳，黎若楠，刘颖旭．新闻数据可视化工具“词云”在新媒体中的创新实践[J]. 中国传媒科技，2021（09）：46-49.

导语

词云图也叫文字云，可将文本中出现的能够代表主要信息的“关键词”予以视觉化的展示。主要流程是通过算法或统计手段获取文字中关键信息，并将这些信息通过形式多样的展示形式加以呈现。这种方式使读者对文本内容有更加直观的体验，是编辑记者最常用的新闻数据可视化形式之一。市面上不乏一些功能和形态较为完善的词云工具，但运用到新闻报道场景中，一方面，词云生成效果较难满足要求，另一方面，容易引发安全和版权等问题。基于此，根据新华社新闻报道的特点，充分调研编辑记者对词云工具的需求，自主研究关键词提取技术和数据可视化技术，研发了一款操作简单、风格明快的在线词云制作工具。在该词云工具的开发和迭代过程中，在产品形态、总体技术架构、新闻关键词提取算法等方面进行了深入的探索和实践。

1. 产品形态

词云工具是一款面向编辑记者的高效、易用的数据可视化工具，同时提供文本分析和词云制作功能，让没有数据分析和平面设计基础的用户也能轻松设计出精美的可视化词云图。如图 1 所示。



图 1 词云工具界面展示

词云工具的制作界面十分简洁，分为菜单栏、左侧展示区和右侧配置区。核心功能分布在右侧配置区，包括数据编辑和图表设置两大块。数据编辑区提供文本分析功能，通过上传文件等方式，进行文本关键词的词频、权重统计等。图表设置区提供词云样式编辑功能，用于自定义词云形状、字体、颜色等。

1.1 产品功能

1.1.1 数据编辑

用户通过上传文件、输入正文等方式，选择词频、权重等抽取类型，工具将对其进行自动文本分析，抽取关键词，迅速在左侧展示区生成默认的词云效果。

抽取关键词结果将在数据表格中显示，用户可对表格中的数据进行在线修改、添加或删除。同时支持对处理后的数据集进行导出下载。

关键词自动抽取支持长词、短词的两种词频（文本中关键词出现的次数）算法，以及权重（根据文本上下文语义关系计算得出）算法。

上传文件支持 txt、doc、docx、pdf 等多种格式，还支持直接输入稿件正文或有效网址等多种方式。

1.1.2 图表设置

考虑到不同新闻场景，支持对词云的形状、颜色、字体、动画等进行个性化配置，调整参数后可在左侧展示区实时看到渲染效果，易用且高效。

词云形状提供了丰富的样式，包括表情、形状、数字、体育等多种类型。支持用户上传图片，自定义词云轮廓，有效扩大了词云的使用场景。

词云主题色默认提供严肃、活泼、庄重、柔和、渐变等多种主题色，且支持获取图片颜色像素点作为文字

颜色，用户可根据偏好灵活选择。丰富的主题色提升了词云的可视化效果，令其在新闻稿件中“效果突出”，一目了然。

词云工具还提供多种免费商用字体供用户使用，编辑无需再重新加载或安装字体，即可实时看到字体效果。同时提供个性化配置，用户可根据需求，对字号、轮廓、动画进行个性化调整，对词云的最终效果进行调优，大大提升用户体验。

2. 总体技术架构

系统架构设计过程中，项目组进行了充分的前期调研，与编辑记者进行了多次交流讨论，通过借鉴多家商业产品从而完成了整体的产品设计。考虑到 B/S 架构分布灵活，维护成本低，只要有网络、浏览器，便可以随时随地查询、制作和修改词云，项目组基于前端可视化技术的积累，完成技术选型策略。下面分别从浏览器端和服务器端进行总体技术架构的详细阐述。

2.1 浏览器端

浏览器端采用 React 框架、Ant Design UI 库、G2 可视化引擎等业界最新技术栈，实现简洁、易用、健壮的前端交互式界面。

React 是用于构建用户界面的 JavsScript 库，偏向于更底层的实现逻辑，便于灵活构建自定义组件。由于其采用了 Virtual DOM 设计思想，当页面重新渲染组件时，React 在 Virtual DOM 上通过 diff 算法寻找到要变更的 DOM 节点，再把本次修改作用到浏览器实际的 DOM 节点上，相当于在 JS 和真实 DOM 中间加了缓存，利用 diff 算法减少了真实 DOM 不必要的操作，从而表现了优越的系统性能，在大型系统架构中得到了广泛使用。

Ant Design UI 库是一套开箱即用的高质量 React 组件库，提供了丰富的基础组件，视觉风格简洁美观，覆盖大部分应用开发的场景，结合 React 强大的生态体系形成了一整套前端解决方案，高效率定制开发用户界面，有效构建管理前端项目。

G2 可视化引擎是蚂蚁金服开源的一套图表库，具有高度的易用性和扩展性。[1] 其以数据驱动的高交互可视化图形语法，可灵活绘制出各种各样的图表类型，有效助力可视化分析。本项目中利用 G2 图形语法，底层优化了基于 D3 实现的词云布局算法，从而动态渲染出大量的文字标签。[2]

算法实现原理如下：

初始化关键词的配置参数，对数据进行排序，从权重最大的关键词开始布局；

一个关键词包含四个顶点，通过坐标表示为一个矩形区域。每个关键词在布局时都要通过碰撞检测算法，检测是否与先前已布局好的关键词位置冲突；

若检测到冲突，则会沿着阿基米德螺旋线重新布局该关键词；

若该词不能沿着螺旋线的任何地址被布局，则会轮询展示下一个关键词。

2.2 服务器端

服务器端主要基于 Node.js、MySQL、Redis 等技术，业务逻辑采用 Express 框架进行开发，可快速方便地创建 API 接口服务。MySQL、Redis 等采用集群式数据存储技术，提高了系统的可靠性和性能。

Node.js 是一个事件驱动、非阻塞式 I/O 的 JavaScript 模型，基于 Chrome （V8 引擎）的 Web 应用程序框架。[3]Node.js 还提供了各种丰富的 JavaScript 模块库，极大地简化了使用 Node.js 来扩展 Web 应用程序的开发工作。

Express 是一个基于 Node.js 平台，灵活、便捷的 Web 开发框架，提供了一系列强大的特性帮助快速创建 Web 应用和 Http 工具。其核心特性包括：（1）通过设置中间件来响应 Http 请求；（2）定义路由表用于执行不同的 Http 请求；（3）通过向模板传递参数来动态渲染 Html 页面等。由于 Express 封装了很多功能包，因此在构建大型项目中，通常采用其作为中间服务层框架，用于处理业务逻辑。

MySQL 是一个多线程 SQL 数据库服务器，它能够快速、有效、安全地处理大量的数据。相对于 Oracle 等数据库来说，MySQL 的使用更加简洁，由于其快速、健壮和易用的优点，在 Web 应用方面得到了广泛的使用。

Redis 是一个高性能的数据结构服务器，可以用作数据库、缓存、消息代理。其支持的数据结构包括字符串、哈希、列表、集合、有序集合、位图、超级日志，通过 Redis 哨兵和 Redis 集群自动分区。Redis 运行在内存中也可以持久化到磁盘，具有非常广泛的应用场景，对关系型数据库也能起到很好的补充作用。

本项目中，使用 MySQL 技术保存模板配置，存储用户记录，完成复杂的统计查询操作；使用 Redis 技术存储任务 ID，应用于大体量的文本关键词抽取，通过异步轮询方式，保证了接口的稳定性和可访问性。

3. 关键词提取算法

在词云工具中，采用基于词图模型的无监督方法来提取文本关键词。

文档关键词表征了文档主题性和关键性的内容，是人们快速了解文档内容、把握主题的重要方式。关键词广泛应用于新闻报道、科技论文等领域，方便人们高效地查阅、管理和检索文档。

文档关键词需要同时具可读性、相关性和覆盖度。

可读性：关键词本身应该是有意义的词或者短语。

相关性：关键词必须与文档主题相关。

覆盖度：关键词要能够对文档的主题有较好的覆盖，不能只集中在文档某个主题而忽略了文档其他主题。

文本的关键词提取方法分为有监督、半监督和无监督三种。[4] 有监督的方法将关键词抽取问题转化为每个词的分类问题，对每个词进行是不是关键词的判别，需要进行大量的数据标注。半监督的方法利用少量的训练样本构建关键词抽取模型，然后使用该模型对新的文本数据进行关键词提取，将得到的这些关键词进行人工过滤，将过滤得到的关键词加入训练集，重新训练模型。无监督不需要人工标注数据，利用一些方法发现文本中比较重要的词作为关键词。有监督的关键词提取算法需要高昂的人工成本，半监督需要部分标注数据，同时也需要大量的人工干预，因此现有的文本关键词提取主要采用适用性较强的无监督关键词抽取。

无监督的方法主要有基于统计特征、基于词图模型和基于隐含主题模型三种。基于统计特征的方法，根据词或短语的词性、词频、TF–IDF、位置信息、互信息、词跨度等量化指标进行排序后，选取分值靠前的词或短语作为关键词。基于词图模型的方法，以候选关键词为顶点，以词与词之间的共现关系为边组成一个有向图，然后使用特定的算法来选取出图中比较重要的顶点作为关键词。基于隐含主题模型的方法，利用主题模型中关于主题的分布的性质进行关键词提取。基于词图模型的方法是目前较为常用的方法，本项目的方法在此基础上进行优化。

3.1 新闻关键词提取算法流程

考虑到实际应用中的稿件为新闻稿件，提取的关键词需要为有含义的词或短语，才能对文章有较强的概括性。我们采用以下方法进行关键词提取。

预处理：去除新闻稿件中与稿件内容和结构无关的特殊字符，然后进行分词、词性标注、命名实体识别等处理。

生成关键词候选集，包括三个步骤：

关键短语生成：单独使用分词结果时，由于分词的粒度太细，无法找出来关键短语，这里采用基于依存句法的方式进行关键短语的生成；

根据规则和词表获取新闻稿件中的正向词；

根据规则和指标筛选出合适的关键词候选集；

关键词排序和选择：主要使用两种排序方法，一种是基于权重的排序，一种是基于词频的排序；排序后根据所需关键词个数排序靠前的词作为关键词结果。

算法细节：使用的分词词性标注算法可标注 59 个词性；命名实体识别可识别人名、地名、机构名、会议名和时间词。入选关键词的词性主要有各类名词、动词和实体词；排除关键词的词性有：动词“是”、动词“有”、趋向动词、形式动词、能愿动词、数词。

下面主要描述算法中的两个关键步骤——利用依存句法进行短语提取和关键词排序的详细过程。

3.2 利用依存语法进行短语提取

依存句法关系描述的是句子中各单位成分之间句法层面的关系，也就是各个词语之间的依存关系，这种依存关系不但指出了词语之间在句法上的搭配关系，也可以表明一定的语义关联性。为了尽可能地找到具有实际意义的短语，根据语料归纳和语言学规则整理得出以下 15 种句法依存关系来描述新闻稿件中的依存关系。这 15 种关系具体为：主谓关系（SBV）、动宾关系（VOB）、间宾关系（IOB）、前置宾语（FOB）、兼语（DBL）、定中关系（ATT）、状中结构（ADV）、动补结构（CMP）、并列关系（COO）、介宾关系（POB）、左附加关系（LAD）、右附加关系（RAD）、独立结构（IS）、标点（WP）、核心关系（HED）。

为了保证句法依存的效果能够更好地适应新闻稿件数据，从新闻数据中选择了 1672 个句子，对其进行依存句法的标注后加入到原有训练数据中，提升了句法依存算法对新闻数据的适应性。

在对句子进行词性标注和句法依存分析后，对分析结果进行合并，主要以合并定中结构短语为主，为了防止合并过多，还使用一些规则进行处理，例如：连词、助词和标点等不出现在短语结构中；主谓关系不能合并；部分动补结构可以作为短语等。

表 1 分词与短语提取结果对比表

原句	真金白银的投入，为打赢脱贫攻坚战提供了强大资金保障。
分词结果	真金白银 的 投入 ， 为 打赢 脱贫 攻坚战 提供 了 强大 资金 保障 。
短语提取结果	真金白银的投入 ， 为 打赢 脱贫攻坚战 提供 了 强大资金保障 。

3.3 设置正向词

正向词的加入，对政治类新闻的关键词提取效果有提升作用。

正向词是指对稿件内容具有概括作用的、有积极作用的词语，在某些分类（例如：政治）的新闻中，正向词具有指引作用。

通过和用户沟通，使用两个方式获取正向词，一是获取文本中引号、括号和书名号中的文字作为正向词，二是通过用户设置正向词典的方式。

在某些分类中，正向词的作用十分重要，为了提高正向词的权重，在进行权重排序时，通过加大正向词初始权重的方式提高正向词的权重，以保证其有更大概率出现在关键词排序靠前的位置。

用户反馈证实，正向词提升了政治类新闻关键词的提取效果。

3.4 关键词排序方法

从实际应用场景出发，综合用户意见，提出了以下两种关键词的排序方法：基于权重的排序方法和基于词频的排序方法。

3.4.1 基于权重的方法

得到候选词集后，使用基于 TextRank 的方法对词进行排序。TextRank 算法计算时只依赖于词或短语与其他词或短语的共现度。其步骤如下：

构建图，以词作为顶点，两个词在一定的窗口内共现，则构建边。

应用 PageRank[5] 算法或相似算法获得每个顶点的权重。

基于权重对顶点排序并选择部分词作为关键词。

该方法的优点是，可以基于词语和周围词语的关联程度确定该词的权重，相比 tfidf 等算法，可以通过词语间的相互联系判断该词的重要性。算法中，使用词语的 tf 作为初始值，词窗口选择 5，并且在确定词的初始权重时，加入了正向词的机制，保证正向词的权重。

3.4.2 基于词频的方法

词频指一个词在文章中的出现次数，出现次数越多的词代表该词越重要。基于用户反馈，得知权重排序算法给出的权重值并不能很直观的体现词汇的重要性，因此，在生成候选关键词集后，可以选择基于词频的方法得到关键词。

基于词频的统计方法分为分词词频和出现词频两种。

分词词频：在通过分词、词性标注和命名实体识别等步骤后得到的候选词集上对其进行词频统计，这样统计得到的词频是正确分词后的该词的次数。

出现词频：在得到关键词候选集上，统计每个词在文中的出现次数，使用的是字符串匹配的方式，这种方式比较直观，和常用的查找方式结果一致。

分词词频的统计结果更准确，而出现词频的结果更方便验证，这两种方式可由用户进行自由选择。

此外编辑还提出，算法提取的词汇比较短，并不能看出真实的意思体现，从美观的角度看，有些图比较适合用长词填充，有些词比较适合用短词填充。按照以上需求，又提供了过滤短词的方法。

由系统默认配置和用户指定的方式设置短词长度，小于等于该短词长度的词称为短词，大于该短词长度的词称为长词。如果短词在长词中出现，那么需要在关键词候选集中去除该短词。这样既可以保证长词效果，又可以保证重要的短词不被过滤。在使用过程中这种方式的提取效果也受到了用户的认可。

4. 成果及改进方向

通过前期调研、产品设计、自主研发、算法调优，项目组迅速完成了代码编写和高效率地迭代开发，实现了整个词云工具的构建和优化。基于该工具，记者编辑在 2021 年春节、两会等重大报道期间制作了多个效果出色的作品，并对该工具表示了肯定，同时也提出了宝贵的意见和建议。

接下来会从优化用户体验、提供更灵活的自定义设置、提供更多更精美的样式效果等方面进行改进，通过不断打磨，让产品继续成长。

5. 总结

通过自主研发的方式，针对新闻场景进行创新应用，结合业务的分析和改进，最终形成了具备新闻报道要求的在线词云工具，有效助力编辑快速创作数据新闻。以此为基础，将探索更多新闻数据可视化工具的研发。媒

参考文献

[1] Leland Wilkinson.The Grammar of Graphics[M].Springer，2011.

[2] Jonathan Feinberg.Beautiful Visualization[M].O’Reilly Media，2010：37-58.

[3] 朴灵 . 深入浅出 Node.js[M]. 北京：人民邮电出版社，2013.

[4] Hasan K S ，Ng V. Automatic Keyphrase Extraction： A Survey of the State of the Art[C]. Meeting of the Association for Computational Linguistics. 2011.

[5] Page，L & Brin，S & Motwani，R & Winograd，T. The PageRank citation ranking： Bringing order to the Web. Technical report.

作者简介：秦玉芳（1987-），女，河南焦作，工程师，研究方向：数据新闻；黎若楠（1993-），女，宁夏贺兰，工程师，研究方向：数据新闻；刘颖旭（1988-），女，陕西榆林，工程师，研究方向：数据新闻。

（责任编辑：张晓婧）